E
Distr.
LIMITED
ST/ECA/SEM/8
5 July 1981
Original: ENGLISH

UNITED NATIONS

ECONOMIC AND SOCIAL COUNCIL

ECONOMIC COMMISSION FOR AFRICA

Seminar on Household Surveys

Addis Ababa, 21-30 September 1981

PROBLEMS OF SURVEY DESIGN

C O N T E N T S

		Paragraph	Page
I.	INTRODUCTION	1	1
II.	HOUSEHOLD AND HOUSEHOLD SIZE	2 - 9	1 - 2
III.	SAMPLING UNITS AND FRAME	10 - 20	2 - 4
IV.	STRATIFICATION	21 - 42	4 - 8
V.	SURVEY DESIGNS	43 - 96	8 - 17
VI.	MASTER SAMPLE	97 -109	17 - 19
VII.	CONCLUSION	110 -113	20

I. INTRODUCTION

1. The National Household Survey Capability Programme (NHSCP) of which the African Household Survey Capability Programme (AHSCP) is a regional component envisages a continuous and co-ordinated series of national household surveys to generate integrated data over time on a wide range of subjects. NHSCP is aimed to be cost-effective as the unit cost of an individual survey in a continuing programme will decrease and the quality of data will also increase in the long run. To satisfy the objectives of NHSCP, the survey design plays an important role.

II. HOUSEHOLD AND HOUSEHOLD SIZE

2. By the United Nations[1/] Principles and Recommendations for Population and Housing Censuses, a household may be either:

(a) a one-person household, that is, a person who makes provision for his or her own food or other essentials for living without combining with any other person to form part of a multi-person household, or

(b) a multi-person household, that is, a group of two or more persons living together who make common provision for food or other essentials for living.

The persons in the group may pool their income and have a common budget to a greater or lesser extent; they may be related or unrelated persons, or a combination of both.

3. It further defines households in terms of living quarters: "Households usually occupy the whole, part of or more than one housing unit but they may also be found living in camps, boarding houses or hotels or as administrative personnel in institutions, or they may be homeless".

4. The above definition, although broad, will not cover all types of African households with complex living and eating arrangements among husbands, wives, children and relatives. The Economic Commission for Africa[2/] defined the household as follows: "A household is defined as a person or group of persons who live and eat together". This basic concept has proved a difficult one for many African countries. Attempts to adapt it to suit local conditions have met with varying degrees of success.

5. One such attempt to adapt the concept of the household as a socio-economic unit to the conditions of a country was made in the 1970 census of Ghana. The definition used is quoted in full below to illustrate the type of adaptation which African countries can make to the general concept. The definition was as follows:

> "A household consists of a person or a group of persons who live together in the same house or compound, share the same house-keeping arrangements and are catered for as one unit. It is important to remember that members of a household are not necessarily related (by blood or marriage) because maid-servants may form part of a household. On the other hand, not all those related in the same house or compound are necessarily members of the same household. Two brothers who live in the same house with their wives and children may or may not form separate households depending on their catering arrangements. The same can be said of a father and his married children. Thus in many cases a house or compound may be broken up into separate households".

1/ United Nations Principles and Recommendations for Population and Housing Censuses, New York, 1980 (Sales No. E.80. Vol. 8).

2/ "Household data requirements" E/CN.14/SM/22, Working Group on Organization, Content and Methodology of Household Surveys, Addis Ababa, 15-19 October 1979.

The examples are provided as guidelines in deciding who should form a household in the ECA document referred to above. It should be stressed that, whatever definition of household is used, it should preserve the convenience of recording and ensure complete coverage.

6. Another problem is whether to use a de jure approach or a de facto approach in defining a household size. By the de jure approach a household will consist of all household members who are normally residents of that household (within specified time limits) regardless of whether they are present at the time of data collection. By the de facto approach, a household will cover all household members and the visitors or guests present at the time of data collection.

7. When the de jure approach is used, in surveys with repeated visits, household size will be relatively stable and will have a realistic relation with income. However, this approach may not be able to show the **per capita consumption** of food.

8. When the de facto approach is used in the surveys with repeated visits, household size will vary from one visit to another which can be recorded by it. The de facto approach will be more suitable in evaluating actual consumption of food. It appears that both approaches have advantages and disadvantages.

9. The Handbook of Household Surveys[3/] proposes "The de jure procedure is generally considered preferable in a continuing survey programme, especially where the enumeration is spread over a considerable period of time. Temporary movements of the population in such a situation are less likely to result in omissions or duplications than they would under a de facto approach. In addition, certain kinds of household information such as household size and composition, household incomes and expenditures, etc., are more meaningful analytically under a de jure procedure, since the data reflect the more permanent or usual situation. The de facto approach is considered most suitable when the enumeration is compressed into a very short period or highly transient groups such as homeless persons or nomads".

III. SAMPLING UNITS AND FRAME

10. At the planning stage before formulating a sample survey design, it is necessary to investigate the availability of frames in the country. A sampling frame consists of a list of sampling units or maps of area sampling units covering the whole country or place to be surveyed with proper identification and demarcation.

3/ United Nations, "Handbook of Household Surveys"(revised draft), Introduction and part one, General survey planning and applications, 1981, p 34.

11. In general, the sampling units in household surveys are of a hierarchical nature starting with households as the elementary or ultimate stage units, segments or clusters of households as sampling units in the next stage and census enumeration areas or administrative areas as sampling units at higher stages. The elementary sampling unit is part of a segment which again is part of a higher stage sampling unit and so on. Some surveys use sampling units of more or less stages depending on the available resources and objectives. The types of units will also depend on type of survey, eg., holdings or farms will be used as sampling units in agricultural surveys. The set of sampling units covering all the areas of focus is called the population represented by the sampling frame.

12. Almost all countries have lists of administrative areas such as districts, towns and villages. Maps of these areas defined or definable on the ground are usually available but some are probably out of date. Similarly if a population census has been based on maps of enumeration areas (EAs) with proper boundaries, this information should be available. For these areas, population, number of households or dwelling units and related information may be available through census or administrative records. But if there is a time gap between the last census and the surveys to be initiated, then the maps and information will need updating.

13. Since the frame is essential to prepare a proper survey design, it is highly desirable to update the available frame at the planning stage to ensure that the maps or lists in the frame are accurate and identifiable, covering the whole survey area with no overlapping or omission. However, it would be too expensive to prepare a frame of elementary sampling units such as households over a large area. The work would take too long and turn out to be full of defects.

14. If the survey design involves multistage sampling, the frame of sampling units can be set up in stages. The first frame will cover only the primary sampling units in the first stage and the frames in the second stage will consist of only the secondary sampling units belonging to the selected primary units and so on. In this way cost of establishing the frames will be reduced significantly and moreover these frames will be more reliable. This is one of the reasons for applying multistage sampling in many surveys.

15. In countries where recent population censuses have been based on EA maps, a geographical frame with supporting information should be readily available after some up-dating. However, this is by no means the standard situation as the following examples will illustrate.

16. One African country had taken a successful census with good EA maps five years before planning the first round of its survey programme. All that was necessary was to revise some urban EAs to take into account road construction and new buildings and make corresponding modifications to the related estimates of population and households.

17. A few cases have been found where lack of a previous census or absence of proper EA mapping has made it necessary to start with administrative areas in constructing a frame. In one of these countries intensive work in establishing urban and rural community associations has resulted in a very satisfactory sampling frame. In another the largely unmapped census EAs were invalidated by a change in local government structure; the new local units provided a satisfactory sampling frame although it was difficult to obtain reasonable estimates of their populations and numbers of households.

18. At the other end of the spectrum are cases where the original census EAs have remained largely undisturbed but there have been changes in the administrative districts used as primary units. In countries where this has happened there has been no serious difficulty in producing an effective sampling frame.

19. In general it can be concluded that the establishment of geographical sampling frames in African countries is not a particularly difficult process, but there are no set rules for the exercise because the situation in individual countries has to be taken fully into account. The one common factor is that the work is much easier in countries where a careful census mapping operation has been undertaken. This is why many countries now believe that it is worthwhile keeping EA maps continuously updated.

20. A special problem arises in enumerating nomadic populations becuase they are only partially related to the geographical frames discussed above. Nomads normally require separate sampling arrangements and there are many methods out of which two approaches based on hierarchical lists of chiefs and sub-chiefs and lists of water-points have been used widely in Africa. The hierarchical approach alone is not satisfactory because the records maintained by chiefs are designed to deal with sensitive matters such as taxation. However the seasonal movement of nomadic groups tends to be reasonably stable and the watering points help to locate them during the dry season. A suitable combination of the two approaches can produce an effective sampling scheme. In each approach the frames of sampling units can be prepared in stages.

IV. STRATIFICATION

21. The designs for sample surveys and especially national surveys are preceded by stratification of the population. Yates[4] explained the purposes of stratification as follows.

> "Stratification has two purposes. The first is to increase the accuracy of the overall population estimates. The second is to ensure that subdivisions of the population which are themselves of interest are adequately represented. Such divisions may be termed domains of study. Maximum over-all accuracy will be attained if the strata are so chosen that the units within each stratum are as similar as possible."

22. It means that stratification of the population is a necessary step taken before the selection of a sample in order to improve the accuracy of over-all estimates and at the same time to get separate estimates for each subdivision designated as a domain of study. The over-all sample will be composed of the samples selected independently from each stratum.

23. Some important points to be observed in stratifying the population are that the units within each stratum are as similar or uniform as possible and on the other hand those belonging to different strata are as dissimilar as possible. If this principle is not strictly adhered to, then the estimates in stratified sampling can be found to be less efficient than estimates in direct sampling.

24. Stratification can be carried out using the available supplementary information related to the subjects to be covered in the surveys and the strata will cover the whole area with no overlapping and omission. It means that all sampling units will be included in one or another stratum but no unit will be allowed to be present in more than one stratum. Since all sampling units in the different stages are featured in the frames of those stages in the form of a list or maps, strata will also appear as groups of sampling units in those frames.

4/ Yates, F., 1960, "Sampling Methods for censuses and surveys" Charles Griffin and Co., London, 24.

25. With respect to sample size and estimation of sampling error, Yates further wrote, "Since in a stratified sample only variation within strata gives rise to sampling error, it is this component of stratification that requires estimation, and this can in general only be done from differences between units in the same stratum. It is therefore necessary if an estimate of sampling error is required, that the strata be of such size that the sample contains two or more units from at least the majority of strata".

26. On the question of the number of strata, many statisticians have made studies using the assumed distributions and actual data. Cochran (1977)5/ showed with examples that generally little reduction in variance is to be expected beyond the number of strata = 6 when over-all estimates are to be made. If the estimates are wanted also for geographical subdivisions (presumably domains of study), the argument for a larger number of strata is stronger. Dalenius (1953)6/ showed by an empiral example that the reduction in variance after a certain number of strata is marginal unless additional stratification characteristics are available and used. Kish (1967)7/ agreed with the above authors but added that perhaps between three and ten strata suffice for a single variable. There are however other statisticians like Sethi (1963)8/ who showed that under gamma distribution the efficiency of stratified sampling with proportional allocation increases in direct proportion to a quadratic function of the number of strata.

27. In the integrated survey programme, many subjects will be covered, for which the population can be divided into strata by multiple stratification in either the first stage or later stages. But the number of strata in the first stage will be limited by sample size since at least two units should be selected out of each stratum except in unavoidable cases. It has also been indicated that formation of many strata will not necessarily lead to appreciable increase in the efficiency of estimates. Therefore a few characteristics related to the main subjects should be used for stratification in the first or later stages.

28. In practice, when multistage sampling is to be applied, the population may not need to be divided into many strata in the first stage. It can be divided into regions and urban and rural areas. In each region or all over the country the urban areas can further be divided by size of population and by economic importance, and rural areas are stratified by geographical or agricultural conditions. The strata so formed will most probably constitute domains of study.

5/ Cochran, W.G., Sampling Techniques, third edition, Wiley, (1977) Sec. 5A8.

6/ Dalenius, T., Multivariat Sampling Problem, Skad. Akt. (1953) 36, 92-122.

7/ Kish, L., Survey Sampling, Wiley, (1967) Sec. 3.6.

8/ Sethi, V.K., A note on optimum stratification of population for estimating the population means. Australian Journal of Statistics, (1963) 5, 20-33.

29. In the later stages of a survey sub-stratification can be carried out in more detail using the supplementary information collected in the course of forming sub-units. It means that stratification should be planned and carried out at suitable stages of sampling. The other problems encountered in stratification are demarcation of boundaries between strata and allocation of sample into strata.

30. When the units are to be stratified using a quantitative variable such as population size or crop production, it will be necessary to study the frequency distribution of such a variable and fix the number of strata to be formed. One method due to Delenius and Hodges (1959)[9] was to find the square root of the frequency and create the equal intervals within the cumulation of such roots. The boundaries corresponding to such intervals will provide the boundaries of the strata. This method is applicable when there is a fairly large number of units to be stratified to get a proper distribution of units.

31. There are other methods suggested by several statisticians for demarcating strata boundaries. Sethi[10] tabulated the optimum boundaries for normal and various χ^2 distributions under different allocations of a sample when the number of strata is less than seven. One drawback of this method is that the number of units to be stratified must be large in order to approximate the above distributions.

32. In practice the boundaries of strata in the first stage are fixed in such a way that the values of variables in a stratum are similar generally but as proximate as possible. It means that stratification in the first stage will not be very thorough or detailed. However in the later stages, more refined stratification can be introduced using supplementary information closely related to the major subjects to be surveyed. Experiences in stratification in some developing countries are as follows.

33. In an income and expenditure survey of an African country the whole area was first divided into urban and rural strata. The households in the penultimate stages were listed and visited to collect some basic information including income of individual household members and household income from agriculture and other enterpreneurial activities using special forms. The enumeration records were sent to the headquarters where the households were divided into three sub-strata separately for the urban and rural strata. The sub-stratum boundaries for the urban and rural areas were fixed so as to allocate equal share of total cash income to each sub-stratum and households were selected. This is a case of two-phase sampling applied in the penultimate stage.

34. In another country for the integrated survey, the rural area is first divided into ecological zones and each selected sub-district (arrondissement) is divided again into rural and urban areas. The households in the selected segments in the penultimate stage are listed and their income and the related variables such as ownership of land, livestock and other assets are recorded. Then the households are stratified or arranged in order of values of such supplementary information and some are selected. This is a two-phase sampling.

9/ Delenius, T. and Hodges, J.L. Jr., "Minimum variance stratification"Journal of the Amer. Stat. Assoc., 1959, 54, p 88-100.

10/Sethi, V.K., op. cit.

35. In one country[11] to conduct a [illegible] survey in a rural area, the villages were first stratified on the basis of their location. The selected villages were visited and all available data including the occupation of members were collected from all the households which were then substratified by main occupations (i.e., occupation with highest income of the household) such as farmers, tenants or farm labourers, government employees, animal husbandry or fishing and others.

36. For the rural income distribution survey of an African country,[12] the rural population was divided into two non-overlapping sampling frames, one consisting of employees of freehold farmers whose annual incomes were available from official records and the others whose incomes were not available. To survey the second group, a division was first made between more and less densely populated areas. All the households in the penultimate stage were interviewed, listed and sub-stratified. The first substratum consisted of households with 30 or more head of cattle, the second substratum consisted of households with fewer than 30 head of cattle and with no additional dwelling place outside the village. The third substratum contained the remaining households.

37. In an African multipurpose sample survey[13] the rural areas were stratified into sub-divisions and then the households in the selected sub-divisions were sub-stratified by household class, i.e., owner, tenant, etc.

38. The survey of rural small-scale industrial establishments was carried out in another country[14] by stratifying the villages according to rough measures of the volume of their industrial activity as reported in the agricultural census in the preceding year. The establishments reported were substratified by type of industry such as textile and other activities.

39. Generally the frame of large and medium size industries is available and these are usually stratified by type and size of either production or employment or other important characteristics. Small-scale industries are not properly registered and, even if there is a list, it is subject to frequent changes. Therefore the areas are usually stratified by industrial activities based on the available information and the selected areas are substratified by type and size of industries after listing them.

11/ Myint Tin and Than Toe, "Estimation for domains in multistage sampling", Journal of the Amer. Stat. Assoc., 1972, 67, p 913-916.

12/ United Nations, "Sample surveys of current interest", thirteenth report, New York, 1979, ST/ESA/STAT/SER.C/14.

13/ Ibid.

14/ Ibid.

40. Another issue is allocation of a sample into different strata once the total sample size to be selected is fixed. The common method used is proportional allocation by which the sample is allocated in proportion to size of strata. The second method is optimum allocation. The second method is more effective in minimizing sampling error at a given cost of survey or minimizing cost at a specified limit of sampling error. To achieve these, a larger sample will be chosen from a stratum (a) if stratum size is large; (b) if variability of the main characteristics is high in that stratum; and (c) if the cost of sampling of a unit including cost of travel between units in that stratum is low.

41. In integrated surveys stratified sampling with optimum allocation can be applied in a judicious way. Since many subjects will be covered in each survey cost as well as variance will change. The method of allocation becomes more complicated if multistage and multiphase samplings are applied, for which cost and variation or intracluster correlation at every stage and every phase will have to be taken into consideration. Another point to be observed in relation to allocation is the amount of data on a subject expected to be obtained from each stratum when it is to be treated as a domain of study. This point is important when events investigated are likely to be small in certain areas. But pilot surveys and previous surveys can furnish the necessary information for optimum allocation and hence serious efforts should be made to reduce survey cost at a desirable level of precision by improving the design after each survey. This will make integrated surveys cost-effective, which is a major aim of NHSCP.

42. In fact the experience of many African countries shows that stratification is not really very difficult and most of them do stratification in the first stage (by area) and the penultimate stage (using supplementary information).

V. SURVEY DESIGNS

43. The objective of the sample survey design is to obtain the most reliable estimates at a specified cost or to obtain the estimates with specified level of precision at a minimum cost.

44. The reliability of estimates are assessed by sampling and non-sampling errors. Sampling error is measured by the expected value of the square of the difference between an estimate and the corresponding population value. It is equivalent to the variance of estimate and the square of bias if it exists. Therefore, sampling error is present in all sample surveys at least in the form of variance of estimate whenever sample size is less than the population size. However it can be measured only if the sample is selected at random by assigning probability of being selected to all sampling units.

45. On the other hand non-sampling errors occur in executing censuses or surveys at various stages such as planning, data collection, tabulation and printing. Errors at the planning stage arise from a defective frame and concepts and definitions in the questionnaires. Those in data collection appear in the form of coverage and content errors, owing to faults and failures in organization, field staff and respondents. Errors in tabulation and printing arise in editing, data preparation, processing and printing.

46. The cost of a survey includes expenditure on preparation, data collection and processing. Hence the cost consists of:

(a) overhead costs;
(b) cost of survey of a unit in each stratum and stage; and
(c) data processing.

47. Sample designs applied in most surveys are stratified multistage sampling with possible sub-stratification. Strata or sub-strata may represent domains of study or just mere divisions to make design efficient by covering all sub-divisions in the sample.

48. Some reasons for multistage sampling have been given in section III, the major ones being to:

(a) Save costs in constructing or updating the frame;
(b) Save cost and time of travelling between the penultimate units and households in them;
(c) Control non-sampling errors by better supervision and organization of field staff who are assigned to the small areas of the selected primary sampling units.

49. On the other hand considerable time is needed to design surveys and list and select units in various stages. Care should be taken at every step to have the work done properly. Multistage sampling is however less efficient than direct sampling and efficiency will generally decrease as the number of stages is increased. But cost of constructing the frames in stages is smaller, and since the households and penultimate units such as clusters of households are confined to some areas only, cost and time of travelling between those units is considerably reduced compared to cost of direct sampling of the same number of households. Another advantage of multistage sampling is that supervisors can manage the enumerators assigned to the units in an administrative unit much more efficiently and frequently so that some aspects of non-sampling errors can be controlled. That may be more difficult when the units and enumerators are spread out in direct sampling.

50. Samples in each stratum in each stage can be selected with equal or unequal probability depending on the sizes of units and available information and resources.

Determination of optimum sample sizes

51. Before discussing survey designs in detail, it will be appropriate to determine optimum sample sizes in a general manner. To simplify the problem, it will be assumed that a stratified two-stage sampling is applied in a country as follows.

52. The country is divided into L strata, N is the number of primary sampling units (PSU) in a stratum, M_i is the number of secondary sampling units (SSU) in the i^{th} PSU, and n and m_i are selected out of N and M_i by probability sampling.

53. Then total error consisting of sampling and non-sampling errors is assumed to be of the form

$$V = \sum^{L}\left(\frac{V_1}{n} + \frac{V_2}{n\bar{m}} + B^2 + n E_1 + n\bar{m} E_2\right) \quad - \quad (1)$$

with summation over all strata and finite population correction ignored.

where $\bar{m}$ = average size of m_i

V_1 = first-stage variation in a stratum

V_2 = average second-stage variation

B = bias which may be nil or a value of any order (i.e., bias in estimate and part of sampling error)

E_1 = non-sampling error of a PSU

E_2 = average non-sampling error of a SSU

Total cost of survey can be described by

$$C = c_o + \sum^{L}(nc_1 + n\bar{m} c_2) \quad - \quad (2)$$

where c_o = overhead cost

c_1 = cost of travelling between PSUs and listing and enumerating all SSUs in a PSU of a stratum

c_2 = cost of survey and processing data of a SSU

54. The optimum values of n and $\bar{m}$ are obtained by minimizing the total error at a specified cost or minimizing the cost at a specified total error as:

$$n = \sqrt{\frac{V_1}{Gc_1 + E_1}} \quad - \quad (3)$$

$$\bar{m} = \sqrt{\frac{V_2}{V_1}} \quad \sqrt{\frac{Gc_1 + E_1}{Gc_2 + E_2}} \quad - \quad (4)$$

where G is a constant. In finding these relations, bias has been ignored.

55. Equation 3 shows that the number of PSUs to be selected from a stratum will be:

(a) directly proportional to the first-stage variation within that stratum, and
(b) inversely proportional to cost per PSU and non-sampling error in the first stage in that stratum.

56. Equation 4 indicates that the average number of SSUs to be selected out of a PSU will be

(a) directly proportional to the second-stage variation, cost of a PSU and first-stage non-sampling error in that stratum;
(b) inversely proportional to the first-stage variation, cost of a SSU and second-stage non-sampling error in a PSU of that stratum.

57. One important point is that since the first-stage variation V_1 is directly related to intracluster correlation of a variable between the SSU of a PSU in that stratum, if the intracluster correlation is high, more PSUs but fewer SSUs will need to be selected.

58. If the non-sampling errors E_1 and E_2 are ignored in determining the sample sizes, then the optimum values of n and $\bar{m}$ will become:

$$n = \frac{(C - c_o)\sqrt{\frac{V_1}{c_1}}}{\sum_{L} (V_1 c_1 + V_2 c_2)} \qquad - (5)$$

$$\bar{m} = \frac{\sqrt{V_2/c_2}}{\sqrt{V_1/c_1}} = \sqrt{\frac{V_2}{V_1}} \sqrt{\frac{c_1}{c_2}} \qquad - (6)$$

59. Equation 5 shows that the optimum number of PSUs to be selected from a stratum will be directly proportional to the square root of variation relative to cost of a PSU in that stratum. In other words the optimum value of n will increase if intracluster correlation is high and cost of a PSU is small.

60. Equation 6 points out that the optimum number of SSU to be selected from a PSU of a stratum will be the square root of ratio of relative variation to cost in the second stage to relative variation to cost in the first stage. It means that $\bar{m}$ will decrease as n is increased and vice versa.

61. To illustrate the determination of values of n and $\bar{m}$ it should be assumed that the units in both stages are selected by simple random sampling and the number of PSUs in the population and the number of SSUs in each PSU are large in order to overlook the finite population corrections and to make the other approximations.

62. Then the ratio of V_2 and V_1 can be approximated by

$$\frac{V_2}{V_1} = \frac{1 - r}{r} \qquad - (7)$$

where r is intracluster correlation in that stratum. Using this relationship, a table of $\bar{m}$ for different values of c_1/c_2 and r is obtained as follows:

Table 1 - Approximate values of $\bar{m}$

r \ c_1/c_2	2	4	8	16	32	64
.1	5	6	9	12	17	24
.2	4	4	6	8	12	16
.3	3	3	5	7	9	13
.4	2	3	4	5	7	10
.5	2	2	3	4	6	8

63. For any value of r, the minimum size of $\bar{m}$ should be at least two. The above table shows that, as the intracluster correlation is increased, the number of SSUs to be selected from it will decrease. On the other hand if the cost of a PSU compared to that of a SSU rises, the number of SSUs to be selected will go up.

64. The optimum value of n can be determined by substituting the necessary values in equation 5. An easier approach is to use a 10 per cent relative margin of error with a confidence probability of 95 per cent assuming a large sample, so that:

$$\frac{s}{b\sqrt{n}} \sqrt{r + \frac{1 - r}{\bar{m}}} \leq 5\% \qquad - \quad (8)$$

where b is a value to be estimated and s is the standard deviation of SSUs in the population. Using this relation the following table of n for various values of r and $\bar{m}$ and for $s/_b$ = 20 per cent is obtained.

Table 2 — Approximate values of n (for $s/_b$ = 20 per cent)

r \ $\bar{m}$	5	10	15	20	25
.1	5	3	3	3	3
.2	6	5	4	4	4
.3	7	6	6	6	6
.4	9	8	7	7	7
.5	10	9	9	9	8

65. These tables show that, as intracluster correlation increases more PSUs but fewer SSUs are to be selected. However, if the number of SSUs ($\bar{m}$) is not less than 10, then almost the same number of PSUs need to be selected for each value of intracluster correlation.

66. These tables are intended for illustration only and for a survey the optimum values of $\bar{m}$ and n should be calculated using the appropriate parameters and approach depending on the main subjects or characteristics of the survey.

Size of clusters

67. In finding the optimum values of $\bar{m}$ and hence n, it will be observed that the intracluster correlation r plays a crucial role. Generally the smaller the size of the cluster the larger will be the value of r and vice versa, since units in a smaller cluster are more likely to be similar to each other than those in the larger cluster. However cost will be high if a larger number of small clusters are to be covered. But the number of clusters to be selected is influenced more by r than $\bar{m}$ for a certain value of $s/_b$ as seen in table 2.

68. On the other hand the precision of an estimate in multistage sampling depends to a large extent on the number of PSUs selected in the sample, since variation between PSUs is greater than those between smaller units.

69. Another important factor to be taken into consideration is coverage of subjects or characteristics in integrated household surveys and the number of classes in which some characteristics are to be presented. It means that the total number of units selected to represent a domain of study should be large enough to satisfy these requirements.

70. Non-sampling errors should also be taken into account in formulating the design and determining the sizes of clusters and samples. If the clusters are large in size the households selected from them can be spread out so widely that interviewers can commit some non-sampling errors out of fatigue from travelling.

71. A pilot survey using a sample of some PSUs and units in the following stages should be conducted to collect information on cost, variances, intracluster correlation and estimates of major characteristics of the survey to be undertaken. If such information from previous surveys is available, it should be exploited to determine the optimum cluster and sample size taking fully into consideration the objectives of survey and available resources.

72. Alternatively, an experiment using different sizes of clusters may be conducted to calculate relative net precision which is inversely proportional to the variance obtained for fixed cost to find an optimum size of cluster as shown by Cochran[15] (1977) for single-stage cluster sampling. Sukhatme[16] produced a table on number of clusters of varying sizes at two levels of cost and another table on relative standard errors, for agricultural surveys.

15/ Cochran, op.cit

16/ Sukhatme, P.V. "Sampling theory of surveys with applications", Iowa State University Press, 1954, p262-3.

73. If more than two stages are involved in the sample design, the penultimate unit will be a cluster and usually of about equal size. Thus it can be arranged in some two-stage sampling also. In practice the size of such a cluster is about 5 to 10 times the number of households to be selected from it so that the households in the sample will be about 5 to 10 households apart from one another if they are selected by systematic sampling. Since the number of households selected from a cluster is usually between 10 and 20, size of a cluster is around 100.

74. This is a convenient size although it may not be the optimum size of cluster as costs and intracluster correlation have not been taken into account.

Sampling with probability proportional to size

75. In the above illustration, stratified two stage sampling in which units are selected by simple random sampling is discussed. In fact if the sizes of PSUs vary appreciably, it is common practice to select the PSUs with probability proportional to size and to select the households (SSUs) with equal probability.

76. If the survey is designed to select the sample in more than two stages using an administrative unit or a group of enumeration areas as a PSU then it is likely that sizes of these PSUs will vary. In such cases it is usual to apply sampling with probability proportional to size or an auxiliary variable of which can result in greater efficiency provided that

(a) reliable values of sizes (or auxiliary variables) are available; and
(b) it is related to the characteristics to be estimated.

77. One drawback in applying sampling with probability proportional to size is that it is tedious to find the estimate and its precision even if the above two conditions are fulfilled. This problem can be solved if the units in the following stages are selected in such a way that the design becomes self-weighting.

78. For instance if three-stage sampling is applied in a stratum or domain of study by selecting a PSU and a SSU with probability proportional to size or a certain scheme then the design will be made self-weighting in the second stage if the product of their probabilities is constant. To make it self-weighting over all stages, a constant fraction of ultimate stage units can be selected with equal probability. Alternatively, the design can be made self-weighting at the last stage.

79. On this question it has been pointed out in E/CN.14/SM/37[17] that by making a design self-weighting data processing and integration of survey results become easier and the summary or advance tables can be prepared conveniently either in the field of regional offices without using a computer. On the other hand the self-weighting design is not as flexible as a non-self-weighting design in allocating units in the later stages and in distribution of workload to the interviewers. Many self-weighting designs end up as non-self-weighting owing to a defective frame and undercoverage of units. If countries can solve such problems in getting summary or advance tables then they might as well opt for non-self-weighting designs. However many countries find data processing a big constraint and therefore some countries try to develop

17/ "Survey designs of four countries under African Household Survey Capability Programme", E/CN.14/SM/37, ECA, Working Group on Organization, Content and Methodology of Household Survey, Addis Ababa, 29 June - 3 July 1981.

self-weighting designs by making the probability of selection of penultimate units constant and selection of a constant fraction of ultimate units. Integration is also an added attraction for self-weighting design, since it is easier to analyse and integrate data if the units of the same domain are selected with equal probability. However it should be possible to integrate data when non-self-weighting designs are used.

80. Another snag is the allocation of optimum sample size in different stages in a self-weighting design. If the actual sample sizes do not depart significantly from those optimum sizes then a self-weighting design will be satisfactory in that respect.

81. In stratified probability sampling, an estimate of the Harvitz-Thompson type which is shown to be robust by Godambe and Thompson (1977)[18] can be used.

Replicated sampling

82. Sometimes, the sample is divided into groups called replicated sub-samples after selection so that each sub-sample can produce an independent estimate like a jackknife estimate for the domain of study. Thus standard error can be determined easily particularly when the design is non-self-weighting or when ratio or complicated estimates are used.

83. In many cases, sampling is designed before selection in order to have interpenetrating sub-samples by Mahalanobis[19] or replicated samples by Deming.[20] For instance, these can be obtained by selecting two samples of $n/2$ units each to get two replicated samples or by selecting two units out of each $n/2$ strata. In fact many arrangements are possible to get two or more samples.

84. The units can be selected by probability proportional to size if sizes vary significantly and otherwise either by simple random or by systematic sampling.

Reduction of some stages in a survey design

85. Generally a design involving fewer stages will be more efficient, other factors being equal. For instance, assuming a frame of sub-divisions which can be taken as units in a stage, clusters are formed by dividing the selected sub-divisions in the field and selected as units in the following stage. Instead, an alternative method is to assign the number of possible clusters of about equal size in the sub-division in their frame and to select the segments directly with equal probability.

18/ Godambe, V.P. and Thompson, M.E., "Robust near optimal estimation in survey practices". Proceedings of the 41st session, Int. Statistical Institute, volume 4, 1977, New Delhi, p129-146.

19/ Mahalanobis, P.C., "Recent experiments in statistical sampling in the Indian Statistical Institute", Jour. Rug. Stat. Soc. 109, 1946, p 325-370.

20/ Deming, W.E. "Sampling designs in business research", Wiley, 1960.

86. The clusters selected can be located in the respective sub-divisions. Then clusters can be formed in the same way as above and selected clusters are identified. In these two designs, the field work will be done in the same way so that cost and non-sampling error may be about the same. However, although the second design can be more efficient by eliminating one stage of sampling, more than one cluster can be selected from the same sub-division.

87. This is an adaptation of the selection method of Deming[21] but his original method was replication with equal probability.

Repeated surveys

88. In repeated surveys, the sample is usually partially or completely replaced by new units in each round or at equal intervals with a possibility of returning to the old units later. The objective of such a design is mainly to minimise non-sampling errors due to migration or respondents' and interviewers' fatigue in some cases.

89. The sample can be replaced in different ways and at various stages. If the sample is replaced at the first or following stages then the frames at that stage and following stages will need to be updated or revised to select new units to replenish the sample. Hence it will be very expensive and time consuming to replace the sample at the higher level.

90. In one African country, it is proposed to use a replicated systematic sample design with partial replacement in the surveys under AHSCP, by which some households in each round are divided into four replicates of equal numbers of households and two replicates will be replaced every round or year.

91. In another African country, two replicated sub-samples of households will be selected from each penultimate unit by circular systematic sampling and a sub-sample will be replaced every round by a new one selected from the same penultimate unit.

92. Many African countries use quarterly rotating samples in income and expenditure surveys. A proposal has been made to apply the same approach in labour force surveys every quarter.

93. In one African country stratified three-stage sampling was applied in the income and expenditure survey carried out recently. In the urban stratum, some councils were selected in the first stage, two segments were selected from each council in the second stage and households were selected from each segment. Those households in each segment were divided into two sub-segments so that there were four sub-segments in each council. Then each sub-segment was surveyed for a quarter and replaced completely by another in the second quarter and so on.

21/ Deming, W.E., "On simplification of sampling design through replication with equal probabilities and without stages". Jour. Amer. Stat. Assoc., 51, 1956, p 24-53.

94. Kish[22] described a design in which each monthly sample comprised eight equal portions, one brand new. The segments of each portion were in the sample four consecutive months, then out eight months, then in again for four months. He also discussed various estimates used in the rotating sample particularly in relation to the correlation coefficient in successive periods.

Interviewers' bias

95. In order to determine interviewers' bias a sample design such as interpenetrating sub-samples proposed by Mahalanobis[23] can be used. By that method an interviewer is assigned to a sub-sample formed at random to test his bias.

96. Another approach is to allocate the interviewers at random into different clusters in an administrative area or a PSU so that an interviewer is responsible for at least two clusters. An anlysis of variance based on two-way classification can be carried out to test a hypothesis on interviewers' bias. There are other designs of experiments for this purpose.

VI. MASTER SAMPLE

97. A master sample often termed a "master sampling frame" is a sub-population or a sub-frame consisting of units to represent the population and to draw further samples from it as required.

98. Yates[24] wrote "When a number of surveys covering the same population or aggregate of material are likely to be required, it is sometimes advantageous to construct a master sample, from which smaller samples can be drawn as required by means of a sub-sampling scheme. The use of a master sample has a number of advantages. It enables a more accurate, complete and adequate frame to be constructed than could be justified if the frame were only required for a single survey. It simplifies the selection of samples, since in the sub-sampling only the material contained in the master sample has to be subjected to the selection process. It enables supplementary information to be obtained which is of value in improving the accuracy of the various surveys. And it enables surveys on the same material to be so planned that the same units are not selected an excessive number of times for different surveys - a matter of some importance when the information is obtained by response to questionnaires".

99. The advantages of a master sample, in addition to reducing the cost and time of updating and maintaining the frame for successive surveys and improving the accuracy are integration of data from different rounds and use of rotating samples at the lower stage.

22/ Kish, L., op. cit.

23/ Mahalanobis, P.C., op. cit.

24/ Yates, F., op. cit.

100. On the other hand Kish[25] cautioned "Since listings are 'perishable', they become less valuable with time and eventually deteriorate beyond usefulness. Their useful life is limited to some reasonable period which depends on the rate and nature of changes in the population. Within that period, the need for using supplementary procedures, to record additions and other changes, becomes increasingly important with the passage of time. Compact segments, denoting actual areas, automatically reflect changes within their boundaries, and handle these problems better than block listings. Eventually, boundaries, and measures of size of blocks and segments undergo such drastic charges that trying to live with the old lists and maps becomes more troublesome and costly than preparing new ones".

101. It is accepted that a master sample needs to be updated before selecting a new sub-sample and new maps of some areas may be necessary and it is agreed that the master sample cannot last for a long time. But it is hoped to be usable for one cycle of surveys, i.e., about four or five years by accommodating changes in the frame. Then a new master will be prepared for another cycle of surveys if there is a sufficient ground for doing so. If, however, there are enormous changes possibly in some urban areas then perhaps a new master sample may be required for those areas. The approach to be used in constructing a master sample in one African country is described as an illustration.

102. The country is divided into four ecological zones and two cities and each ecological zone is further divided into urban and rural areas for which separate estimates are required. Hence there will be ten domains of study. The population census provided a list of enumeration areas (EAs), their maps, and population and number of households in those EAs.

103. In each ecological zone, an arrondissement (sub-district) will be the first-stage unit and segments of equal size (about 120 households) and households will be second-stage and third-stage units. In the first-stage sample a selected arrondissement is stratified into urban and rural areas and a number of segments is chosen from each stratum by circular systematic sampling to get the second-stage samples. Finally a fixed proportion of households is selected by circular systematic sampling using two independent starting points to get two replicated sub-samples from each segment.

104. The number of units to be selected in each stage is first determined for the smallest zone. Since no prior information on intracluster correlation and variations in each stage and costs is available, it is not possible to determine optimum sample sizes. However, on the basis of the experience of other countries the minimum sample size required is determined for that zone. The sample sizes for the other zones are calculated by raising that sample size by relative population sizes of those zones and the number of segments to be selected is calculated to make the design self-weighting. The main reasons for adopting this design are to reduce costs of transportation, to control some non-sampling errors by concentrating on field teams and rotating the sub-samples and to simplify computation, integration and analysis of data.

25/ Kish, L., op. cit.

105. In a city, the first-stage units are segments of about equal size selected by circular systematic sampling to obtain two replicated samples from each arrondissement. Out of each chosen segment, a fixed fraction of households are selected as two replicated sub-samples by circular systematic sampling using two independent starting points. The number of segments selected from each arrondissement is fixed by proportional allocation and a fixed proportion of households is selected from each segment since the necessary information is not available for better allocations of the sample. The design is self-weighting and suitable for applying rotational sampling.

106. For the formation of segments, a cartographic team will be sent to an EA from which a segment is to be selected. The teams will demarcate the boundary of the EA and list all the households to divide them into segments of about equal size. That team will be controlled and checked by a supervisor to avoid bias. Then a team composed of supervisor and interviewers will visit that EA to check the listing and segmentation and to collect basic information including income and related information of households in the selected segment. This is regarded as a first phase sample and after arranging the households in order of income, two sub-samples of households are selected, which will become the second-phase sample.

107. The master sample will be built up based on the present sample in which 25 out of 132 arrondissements (about 20 per cent) are selected for the present survey from the ecological zones and all arrondissements from two cities. These arrondissements, being well distributed and apparently heterogeneous in some characteristics such as population, crop pattern and other major economic activities, are considered to be fairly representative of the whole country. At the same time these arrondissements have stable boundaries. Therefore they will constitute the first-stage master sample in each ecological zone. The only problem is that if there is a change in urban and rural areas in an arrondissement of the master sample, it needs to be accommodated in the sample. In the second stage the number of segments can be increased or decreased in the list of selected segments depending on the type of survey and resources available. If more segments are to be selected then sizes and boundaries of EAs need to be checked again. At any rate the boundaries of segments should be updated before a new survey is launched. In the urban area since the first-stage master sample comprises the segments, the list and boundaries of these segments will need to be checked and updated more frequently and closely.

108. In the final stage a replicated sub-sample will be replaced by a new one selected by circular systematic sampling, after relisting the households in the selected segment. This will facilitate the integration of surveys and permit control of some aspects of non-sampling error. In the following surveys the sample size at various stages will be adjusted on the basis of the variation, intracluster correlation, non-sampling error and cost.

109. These points constitute merely an outline for the formation of a master sample and the details have to be worked out. Since a master sample is smaller than the total population, it is easier and less expensive to check and update. Basic information on households in the penultimate stage of this sample can be collected and updated from time to time so that a form of data base and a kind of time series can be established and the results of surveys can be linked and analysed in depth.

VII. CONCLUSION

110. There are many problems in survey designs and only some are dealt with in this paper. The objective of sample survey design is to obtain the most reliable estimate at a specified cost or to obtain the estimate with a specified level of precision at a minimum cost. It is not easy to attain this objective when the subjects to be covered or the characteristics to be estimated are numerous under AHSCP and NHSCP. An additional objective of survey design is the integration of data of the same survey and of different surveys.

111. The importance of constructing reliable frames in stages is stressed and problems are posed in the concept of households and household size. Stratified multistage sampling is presented as a possible design for construction of a master sample and to produce estimates for domains of study. Stratification has been described in detail showing that it can be carried out in stages particularly in the primary and penultimate stages. An approach to reduce the number of stages is also described. The optimum sizes of clusters and size of samples at different stages are discussed using two-stage sampling as an illustration. The sampling schemes with partial replacement or rotating samples for repeated surveys to control some aspects of non-sampling errors are also discussed, along with some methods of allocation of interviewers to test their bias.

112. Data processing is one serious bottleneck encountered by many African countries so that a decision needs to be taken as to whether the survey design would be self-weighting or not given its advantages and disadvantages.

113. Since many African countries have carried out censuses and surveys and plan to undertake more of them, the problems in survey design can be overcome. There are common problems as well as specific problems for particular countries and surveys. Countries should therefore develop their survey designs depending on their needs, conditions, experience, available resources and subjects to be covered. This paper may perhaps be of some help in developing such designs.